

Б. А. Путко, Н. И. Федорова

Задача разделения зон значений параметров короткого временного ряда на примере анализа роста показателей онкомаркера

Рассматривается задача интервальной оценки параметра экспоненциального роста короткого временного ряда при наличии дополнительных параметров. Результаты оценивания применяются в задаче прогнозирования течения онкологического заболевания на основании динамики показателей онкомаркера.

Ключевые слова: экспоненциальный рост; значимые и незначимые параметры; интервальные оценки; критерий онкомаркера.

JEL classification: C10; C13; C19.

1. Введение

Часто возникает задача оценивания параметров временного ряда в ситуации, когда число параметров сравнимо с числом наблюдений. Классическим примером здесь является интерпретация динамики медицинских показателей — количество наблюдений при этом, как правило, мало.

В этих случаях такие классические методы, как (нелинейный) метод наименьших квадратов, метод максимального правдоподобия или метод моментов (Кремер, 2007; Кремер, Путко, 2010; Магнус и др., 2007) могут привести к результатам, достаточно далеким от истины.

В настоящей работе предлагается метод построения доверительных интервалов, который можно рассматривать как некоторую модификацию метода моментов.

2. Модель

Рассматривается модель $y_t = f(t, k, \theta) + \varepsilon_t$, $t = 0, 1, \dots, T$, где f — функция, вообще говоря, нелинейная; все параметры делятся на существенные $\theta = (\theta_1, \dots, \theta_s)$, оценки которых требуется получить, и несущественные $k = (k_1, \dots, k_m)$, значения которых интереса не представляют. Общее число наблюдений T мало и сравнимо с $s + m$.

Пусть имеется статистика $Z = (Z_1, \dots, Z_s)$, удовлетворяющая следующим условиям:

- 1) $E(Z_i)$ монотонны по параметрам θ и k в рассматриваемой области их значений;
- 2) $E(Z_i)$ медленно меняются по k и относительно быстро по q ;
- 3) $\text{Var}(Z_i) \leq \Delta$, где Δ — достаточно малое число;
- 4) распределение Z близко к нормальному.

В этом случае система неравенств

$$Z_i - 2\sqrt{\Delta} \leq E(Z_i) \leq Z_i + 2\sqrt{\Delta}, \quad i = 1, \dots, s,$$

может дать достаточно информативные оценки параметров θ .

В данной работе этот подход к оценке параметров рассматривается в приложении к задаче оценивания прогноза онкологического заболевания по результатам динамики показателей онкомаркера.

Временной ряд в этом случае имеет вид:

$$y_t = \alpha + \gamma\theta^t + \varepsilon_t.$$

Здесь α — базовый показатель маркера, не связанный с опухолью. Как правило, это медленно растущая величина, при наличии онкологического заболевания ее можно считать постоянной, $\gamma\theta^t$ — экспоненциально растущая специфическая составляющая. Прогностическую ценность представляет параметр θ , определяющий скорость роста (агрессивность) опухоли.

Предположим, что имеется «нулевое» наблюдение:

$$y_0 = \alpha + \gamma + \varepsilon_0.$$

Соотношение между α и γ определяется «моментом включения», т.е. началом фиксации роста показателей онкомаркера. Как правило, соотношение $k = \alpha/\gamma$ достаточно близко к единице — при больших значениях k опухоль практически не выявляема, малые же значения бывают при уже далеко зашедшем процессе. В модели разумно полагать $0.5 \leq k \leq 2$.

Ошибки регрессии ε_t , $t = 0, 1, \dots, T$, можно считать нормально распределенными, причем $E(\varepsilon_t) = 0$, $\sigma(\varepsilon_t) = p(\alpha + \gamma\theta^t)$ — линейная гетероскедастичность; p — малое число, как правило, $p \leq 0.05$.

Число наблюдений T мало. В работе рассматривается случай $T = 6$ (например, один раз в полгода в течение трех лет для опухоли предстательной железы или один раз в три месяца в течение полутора лет для опухоли молочной железы).

Как правило, фиксируются следующие результаты:

- 1) быстрый рост — удвоение за два периода: $\theta \approx \sqrt{2}$;
- 2) умеренный рост — удвоение за четыре периода: $\theta \approx \sqrt[4]{2}$;
- 3) медленный рост — удвоение за шесть периодов: $\theta \approx \sqrt[6]{2}$.

На основании статистических наблюдений (Gann et al., 1995; Blackledge, Lowery, 1994) можно выделить следующие области значений параметра θ :

- 1) зона быстрого роста $\theta \geq 1.39$;
- 2) зона умеренного роста $1.18 \leq \theta \leq 1.22$;
- 3) зона медленного роста $\theta \leq 1.13$.

Итак, уравнения регрессии имеют вид:

$$y_t = (\alpha + \gamma\theta^t)(1 + p\xi_t), \text{ где } \xi_t \sim N(0,1).$$

Определим величины $v_t = \frac{y_t}{y_{t-1}}$, $u_t = \frac{y_{t-1}}{y_1}$, $t = 1, \dots, T$.

Тогда $v_t = \frac{k + \theta^t}{k + \theta^{t-1}} \eta_t$; $u_t = \frac{k + \theta^{t-1}}{k + \theta^t} \cdot \frac{1}{\eta_t}$, где $\eta_t = \frac{1 + p\xi_t}{1 + p\xi_{t-1}}$ удовлетворяют следующим соотношениям¹:

¹ Вывод формул содержится в Приложении.

$$\begin{aligned}
E(\eta_t) &= E(1/\eta_t) = 1 + p^2 + \delta_1 p^4, \\
\text{Var}(\eta_t) &= \text{Var}(1/\eta_t) = 2p^2 + \delta_2 p^4, \\
\text{Cov}(\eta_t, \eta_{t-1}) &= \text{Cov}(1/\eta_t, 1/\eta_{t-1}) = -p^2 + \delta_3 p^3, \\
\text{Cov}(\eta_t, \eta_s) &= 0, \text{ при } |t - s| \geq 2,
\end{aligned}$$

где δ_i — величины порядка единицы.

Распределение величин η_t и $1/\eta_t$ не является нормальным, однако близко к нему. Их асимметрия и эксцесс — величины порядка p .

3. Результаты

Рассмотрим статистики

$$U = \frac{1}{T} \sum_{t=1}^T u_t \text{ и } V = \frac{1}{T} \sum_{t=1}^T v_t.$$

Их количественные характеристики имеют вид:

$$\begin{aligned}
E(U) &\approx \frac{1}{T} \sum_{t=1}^T \frac{k + \theta^{t-1}}{k + \theta^t}, \\
E(V) &\approx \frac{1}{T} \sum_{t=1}^T \frac{k + \theta^t}{k + \theta^{t-1}}, \\
\text{Var}(U) &\approx \frac{2p^2}{T^2} \left[\sum_{t=1}^T \left(\frac{k + \theta^{t-1}}{k + \theta^t} \right)^2 - \sum_{t=2}^T \frac{k + \theta^{t-2}}{k + \theta^t} \right], \\
\text{Var}(V) &\approx \frac{2p^2}{T^2} \left[\sum_{t=1}^T \left(\frac{k + \theta^t}{k + \theta^{t-1}} \right)^2 - \sum_{t=1}^T \frac{k + \theta^t}{k + \theta^{t-2}} \right].
\end{aligned}$$

По приведенным формулам получены таблицы значений величин $E(U)$ и $E(V)$ в диапазоне $1.12 \leq \theta \leq 1.42$ и $0.5 \leq k \leq 2.0$ (для $T = 6$). Эти величины монотонны в рассматриваемой области, при этом их вариации при каждом фиксированном θ не превосходят 0.065.

Стандартные отклонения величин U монотонно возрастают по k и монотонно убывают по θ (в рассматриваемой области), таким образом,

$$\sigma(U) \leq \sigma(U) \Big|_{\theta=1.12, k=2.0} = 0.225p.$$

Стандартные отклонения V монотонно возрастают по θ и монотонно убывают по k , таким образом,

$$\sigma(V) \leq \sigma(V) \Big|_{\theta=1.42, k=0.5} = 0.314p.$$

В работе (Путко, 2011) исследовалась задача выявления различия между линейным и экспоненциальным ростом короткого временного ряда. Было показано, что достоверное различие достигается при $p = 0.015$. На практике, как правило, имеет место меньшая точность, однако можно считать более или менее достоверной справедливость следующих неравенств:

$$U - 0.01 \leq E(U) \leq U + 0.01,$$

$$V - 0.01 \leq E(V) \leq V + 0.01,$$

которые задают доверительный интервал для параметра θ .

Используя малую вариабельность величин $E(U)$, $E(V)$ по параметру k , можно получить достоверные границы (U_L, U_P) и (V_L, V_P) для статистик U, V при различных значениях θ (табл. 1).

Таблица 1. Границы для статистик U, V при различных параметрах θ .

θ	U_L	U_P	V_L	V_G
1.12	0.910	0.944	1.038	1.097
1.13	0.903	0.940	1.043	1.105
1.14	0.897	0.936	1.047	1.113
1.15	0.891	0.931	1.052	1.121
1.16	0.884	0.927	1.057	1.128
1.17	0.878	0.922	1.063	1.137
1.18	0.871	0.918	1.068	1.145
1.19	0.865	0.913	1.073	1.153
1.20	0.859	0.909	1.078	1.161
1.21	0.853	0.904	1.084	1.169
1.22	0.847	0.900	1.090	1.178
1.23	0.841	0.895	1.095	1.186
1.24	0.835	0.890	1.101	1.194
1.25	0.829	0.886	1.107	1.203
1.26	0.823	0.881	1.113	1.211
1.27	0.817	0.876	1.119	1.220
1.28	0.811	0.872	1.125	1.228
1.29	0.805	0.867	1.131	1.237
1.30	0.800	0.862	1.137	1.245
1.31	0.794	0.858	1.144	1.254
1.32	0.789	0.853	1.150	1.263
1.33	0.783	0.848	1.157	1.271
1.34	0.778	0.843	1.163	1.280

Окончание табл. 1

θ	U_L	U_H	V_L	V_H
1.35	0.772	0.839	1.170	1.289
1.36	0.767	0.834	1.176	1.298
1.37	0.762	0.829	1.183	1.306
1.38	0.757	0.825	1.190	1.315
1.39	0.751	0.820	1.197	1.324
1.40	0.746	0.816	1.204	1.333
1.41	0.741	0.811	1.211	1.342
1.42	0.736	0.806	1.218	1.351

4. Выводы

Получены следующие достоверные границы для статистик U и V :

- 1) медленный рост ($\theta \leq 1.13$) — $U \geq 0.9$, $V \leq 1.105$;
- 2) умеренный рост ($1.18 \leq \theta \leq 1.22$) — $0.84 \leq U \leq 0.92$, $1.06 \leq V \leq 1.18$;
- 3) быстрый рост ($\theta \geq 1.39$) — $U \leq 0.825$, $V \geq 1.19$.

Как видно, имеется зона неопределенности $0.84 \leq U \leq 0.9$, $1.06 \leq V \leq 1.105$, в которой не удастся достоверно различать медленный и умеренный рост. При этом в зоне неопределенности вариабельность θ тем больше, чем больше значение k , т. е. раньше произошло «включение» — фиксация показателей роста онкомаркера. Очевидно, это отражает общую проблему онкологии — трудность прогноза на ранних стадиях.

Однако наиболее важная задача — достоверное различие зоны быстрого и умеренного роста — имеет решение: интервалы значений наблюдаемых статистик для соответствующих зон значений θ не пересекаются. Очевидно, что нулевой гипотезой всегда должно служить предположение именно о быстром росте. Критерии $U \geq 0.83$ или $V \leq 1.19$ позволяют отвергнуть эту гипотезу на уровне значимости 0.05. Отметим, что на практике, как правило, достаточно использовать лишь одну из статистик U или V .

Авторы благодарны Н. Ш. Кремеру и В. А. Романову за полезные обсуждения.

Список литературы

- Кремер Н. Ш. (2007). *Теория вероятностей и математическая статистика*. М.: ЮНИТИ-ДАНА.
- Кремер Н. Ш., Путко Б. А. (2010). *Эконометрика*. М.: ЮНИТИ-ДАНА.
- Магнус Я. Р., Катышев П. К., Пересецкий А. А. (2007). *Эконометрика. Начальный курс*. 8-е изд. М.: Дело.
- Путко Б. А. (2011). Определение характера роста основной тенденции временного ряда при малом количестве наблюдений. *Прикладная эконометрика*, 22 (2), 93–97.
- Gann P. H., Hennekens C. H., Stampfer M. J. (1995). A prospective evaluation of plasma prostate-specific antigen for detection of prostatic cancer. *Journal of the American Medical Association*, 273 (4), 289–294.
- Blackledge G. R. P., Lowery K. (1994). Role of prostate-specific antigen as a predictor of outcome in prostate cancer. *The Prostate*, 5, 34–38.

Приложение

Для вывода практических результатов следует учесть некоторые особенности распределения величин η_t , с учетом того, что они отражают реальные измерения и, следовательно, имеют ограничения на область возможных значений. Поэтому предположение о нормальности распределения величин ξ_t нуждается в определенных уточнениях.

Естественно потребовать выполнения следующих двух условий:

1) величины $1 + p\xi_t$ могут принимать только положительные значения (а фактически значения не меньше 0.5);

2) выборочные наблюдения величин ξ_t на практике можно рассматривать как «почти нормальные».

Этим условиям удовлетворяет следующее допущение: плотность вероятности $\varphi(x)$ величин ξ_t совпадает со стандартным нормальным при $|x| \leq a$ и равна нулю при $|x| \geq a + \delta$, где a — квантиль достаточно близкого к единице порядка, δ — некоторое положительное число. На промежутке $a \leq |x| \leq a + \delta$ плотность вероятности может быть выбрана произвольно так, чтобы общая плотность была бы непрерывной, симметричной и монотонно убывающей при $x \geq 0$ функцией. Из неравенства

$$\int_a^\infty e^{-\frac{x^2}{2}} dx \leq \frac{1}{a} \int_a^\infty x e^{-\frac{x^2}{2}} dx = \frac{1}{a} e^{-\frac{a^2}{2}}$$

следует, что число δ можно выбрать из условия $\delta \leq 1/a$.

В частности, приведенные выше условия будут выполнены, если значение a выбрать из диапазона $6 \leq a \leq 1/(2p)$.

При сделанных допущениях получаем следующие оценки количественных характеристик величин η_t :

$$E(\eta_t) = E\left(\frac{1 + p\xi_t}{1 + p\xi_{t-1}}\right) = E(1 + p\xi_t) E\left(\sum_{k=0}^{\infty} (-1)^k p^k \xi_{t-1}^k\right) = \sum_{k=0}^{\infty} p^{2k} E(\xi_{t-1}^{2k}),$$

$$E(\xi_{t-1}^{2k}) = \int_{-\infty}^{\infty} x^{2k} \varphi(x) dx = E_1(\xi_{t-1}^{2k}) + E_2(\xi_{t-1}^{2k})^+ + E_2(\xi_{t-1}^{2k})^-,$$

$$\text{где } E_1(\xi_{t-1}^{2k}) = \int_{-a}^a x^{2k} \varphi(x) dx; \quad E_2(\xi_{t-1}^{2k})^+ = \int_a^{a+\delta} x^{2k} \varphi(x) dx; \quad E_2(\xi_{t-1}^{2k})^- = \int_{-a-\delta}^{-a} x^{2k} \varphi(x) dx.$$

Имеем $E_1(\xi_{t-1}^{2k}) = \frac{1}{\sqrt{2\pi}} \int_{-a}^a x^{2k} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \Gamma\left(k + \frac{1}{2}\right) G\left(k + \frac{1}{2}, \frac{a^2}{2}\right)$, где $G(t, x)$ — неполная гамма-функция.

Функция $f(x) = x^{2k} e^{-\frac{x^2}{2}}$ возрастает при $x \leq \sqrt{2k}$ и убывает при $x \geq \sqrt{2k}$. Таким образом,

$$E_1(\xi_{t-1}^{2k}) = \sqrt{\frac{2}{\pi}} \int_0^a x^{2k} e^{-\frac{x^2}{2}} dx \leq \int_0^a \sqrt{\frac{2}{\pi}} (\sqrt{2k})^{2k} e^{-k} dx = \sqrt{\frac{2}{\pi}} e^{-k} (\sqrt{2k})^{2k} a \leq \sqrt{\frac{2}{\pi}} e^{-k} a^{2k+1}$$

при $k \leq \frac{a^2}{2}$.

$$E_1(\xi_{t-1}^{2k}) = \sqrt{\frac{2}{\pi}} \int_0^a x^{2k} e^{-\frac{x^2}{2}} dx \leq \sqrt{\frac{2}{\pi}} e^{-\frac{a^2}{2}} a^{2k+1} \quad \text{при } k \geq \frac{a^2}{2}.$$

Следовательно,

$$\begin{aligned}
 E_1(\eta_t) &= 1 + p^2 E_1(\xi_{t-1}^2) + p^4 E_1(\xi_{t-1}^4) + p^6 E_1(\xi_{t-1}^6) + \sum_{k=4}^{\lfloor a^2/2 \rfloor} p^{2k} E_1(\xi_{t-1}^{2k}) + \sum_{k=\lfloor a^2/2 \rfloor+1}^{\infty} p^{2k} E_1(\xi_{t-1}^{2k}) \leq \\
 &\leq 1 + p^2 E_1(\xi_{t-1}^2) + p^4 E_1(\xi_{t-1}^4) + p^6 E_1(\xi_{t-1}^6) + \sum_{k=4}^{\lfloor a^2/2 \rfloor} \sqrt{\frac{2}{\pi}} e^{-k} a^{2k+1} p^{2k} + \sum_{k=\lfloor a^2/2 \rfloor+1}^{\infty} \sqrt{\frac{2}{\pi}} e^{-\frac{a^2}{2}} a^{2k+1} p^{2k} \leq \\
 &\leq 1 + p^2 E_1(\xi_{t-1}^2) + p^4 E_1(\xi_{t-1}^4) + p^6 E_1(\xi_{t-1}^6) + \sqrt{\frac{2}{\pi}} e^{-4} a \sum_{k=4}^{\lfloor a^2/2 \rfloor} (pa)^{2k} + \sqrt{\frac{2}{\pi}} e^{-\frac{a^2}{2}} a \sum_{k=\lfloor a^2/2 \rfloor+1}^{\infty} (pa)^{2k} \leq \\
 &\leq 1 + p^2 E_1(\xi_{t-1}^2) + p^4 E_1(\xi_{t-1}^4) + p^6 E_1(\xi_{t-1}^6) + \sqrt{\frac{2}{\pi}} e^{-4} a \sum_{k=4}^{\infty} (pa)^{2k} = \\
 &= 1 + p^2 E_1(\xi_{t-1}^2) + p^4 E_1(\xi_{t-1}^4) + p^6 E_1(\xi_{t-1}^6) + \sqrt{\frac{2}{\pi}} e^{-4} a \frac{(pa)^8}{1 - (pa)^2} \leq \\
 &\leq 1 + p^2 E_1(\xi_{t-1}^2) + p^4 E_1(\xi_{t-1}^4) + p^6 E_1(\xi_{t-1}^6) + \frac{4}{3} \sqrt{\frac{2}{\pi}} a e^{-4} \left(\frac{1}{2}\right)^8 \leq 1 + p^2 E_1(\xi_{t-1}^2) + p^4 E_1(\xi_{t-1}^6) + 10^{-4}.
 \end{aligned}$$

Соответственно,

$$\begin{aligned}
 \sum_{k=1}^{\infty} p^{2k} \left(E_2(\xi_{t-1}^{2k})^+ + E_2(\xi_{t-1}^{2k})^- \right) &\leq p^2 \left(E_2(\xi_{t-1}^2)^+ + E_2(\xi_{t-1}^2)^- \right) + p^4 \left(E_2(\xi_{t-1}^4)^+ + E_2(\xi_{t-1}^4)^- \right) + \\
 &+ \frac{\delta}{\sqrt{2\pi}} e^{-\frac{a^2}{2}} \sum_{k \geq 3} [p(a + \delta)]^{2k} \leq \\
 &\leq p^2 \left(E_2(\xi_{t-1}^2)^+ + E_2(\xi_{t-1}^2)^- \right) + p^4 \left(E_2(\xi_{t-1}^4)^+ + E_2(\xi_{t-1}^4)^- \right) + \frac{4}{3} \cdot \frac{1}{a\sqrt{2\pi}} e^{-\frac{a^2}{2}} \left(\frac{1}{2}\right)^6 \leq \\
 &\leq p^2 \left(E_2(\xi_{t-1}^2)^+ + E_2(\xi_{t-1}^2)^- \right) + p^4 \left| \left(E_2(\xi_{t-1}^4)^+ + E_2(\xi_{t-1}^4)^- \right) + 10^{-4} \right|.
 \end{aligned}$$

Таким образом, $E(\eta_t) = 1 + p^2 + 3p^4 + s$, где $s < 2 \cdot 10^{-4}$, т.е. $E(\eta_t) = 1 + p^2 + \delta_1 p^4$, где $\delta_1 < 4$. Аналогично можно получить: $\text{Var}(\eta_t) = 2p^2 + \delta_2 p^4$ и $\text{Cov}(\eta_t, \eta_{t-1}) = -p^2 + \delta_3 p^4$.